BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 8: e138477
doi: 10.3897/biss.8.138477

OPEN ACCESS

Conference Abstract

Who is Doing Taxonomy, Whereabouts, and Who Is Funding Them? A Practical Test of What Knowledge Graphs Can Tell Us about Taxonomic Research

Roderic Page ‡

‡ University of Glasgow, Glasgow, United Kingdom

Corresponding author: Roderic Page (roderic.page@glasgow.ac.uk)

Received: 03 Oct 2024 | Published: 04 Oct 2024

Citation: Page R (2024) Who is Doing Taxonomy, Whereabouts, and Who Is Funding Them? A Practical Test of What Knowledge Graphs Can Tell Us about Taxonomic Research. Biodiversity Information Science and Standards 8: e138477. https://doi.org/10.3897/biss.8.138477

Abstract

What is the current state of taxonomy? Quentin Wheeler on his podcast "Species Hall of Fame" fears for taxonomy's future, whereas Lucas Joppa and colleagues have famously argued that we've never had so many taxonomists as we do now (Joppa et al. 2011). There have been global surveys of taxonomic research (Grieneisen et al. 2014) but these rapidly go out of date, limiting their utility. Is there a way to have a “dashboard” that summarises the state of the field in terms of who is doing taxonomy, where it is being done, and who is funding it?

The immediate motivation for this talk comes from a tool I recently developed to track the recent taxonomic literature. Inspired by work by the late David Remsen on uBioRSS (Leary et al. 2007), I created BioRSS (Page 2021), which subscribes to Really Simple Syndication (RSS) feeds for a range of taxonomic journals. Papers listed in these RSS feeds and searches (here referred to as “works”) are aggregated and then tagged by geography and taxonomy, much as envisioned by Mindell et al. (2011). Based on the title and abstract, I attempt to classify the work by geographic and taxonomic scope. Geographic tagging uses the “Glasgow Geoparser,” which uses FlashText search (Singh 2017) to match words in the text to high-level geographic names obtained from

© Page R. This is an open access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Wikidata. Patrick Leary's TaxonFinder is used to locate taxonomic names in the text, these are then matched to the Global Biodiversity Information Facility (GBIF) using the Global Names verifier. For each matched name, the path from taxon to root (the taxon's "lineage") is represented as an array of strings. The majority-rule consensus (Margush and McMorris 1981) of these paths determines what taxon the work is primarily about.

To navigate this data, I created a simple web site that provides a treemap view of the GBIF classification, a map, and a list of works ordered from most recent to oldest (Fig. 1). Exploring BioRSS, one gets a sense of in which countries most new species are discovered, and which taxonomic groups those discoveries fall into. How do we gain more insight into these patterns? One approach, sketched in Page (2023), would be to combine linked data from taxonomic name databases (via Life Science identifiers (LSIDs) for taxonomic names) with data from CrossRef of publications (via Digital Object Identifiers, DOIs) and ORCID (Open Researcher and Contributor ID) on people and their affiliations (via ORCIDs) into a single knowledge graph. By traversing this graph from name to publication to people to institution, we could gain insights into who is publishing taxonomic work, where they are based, and who is funding them.

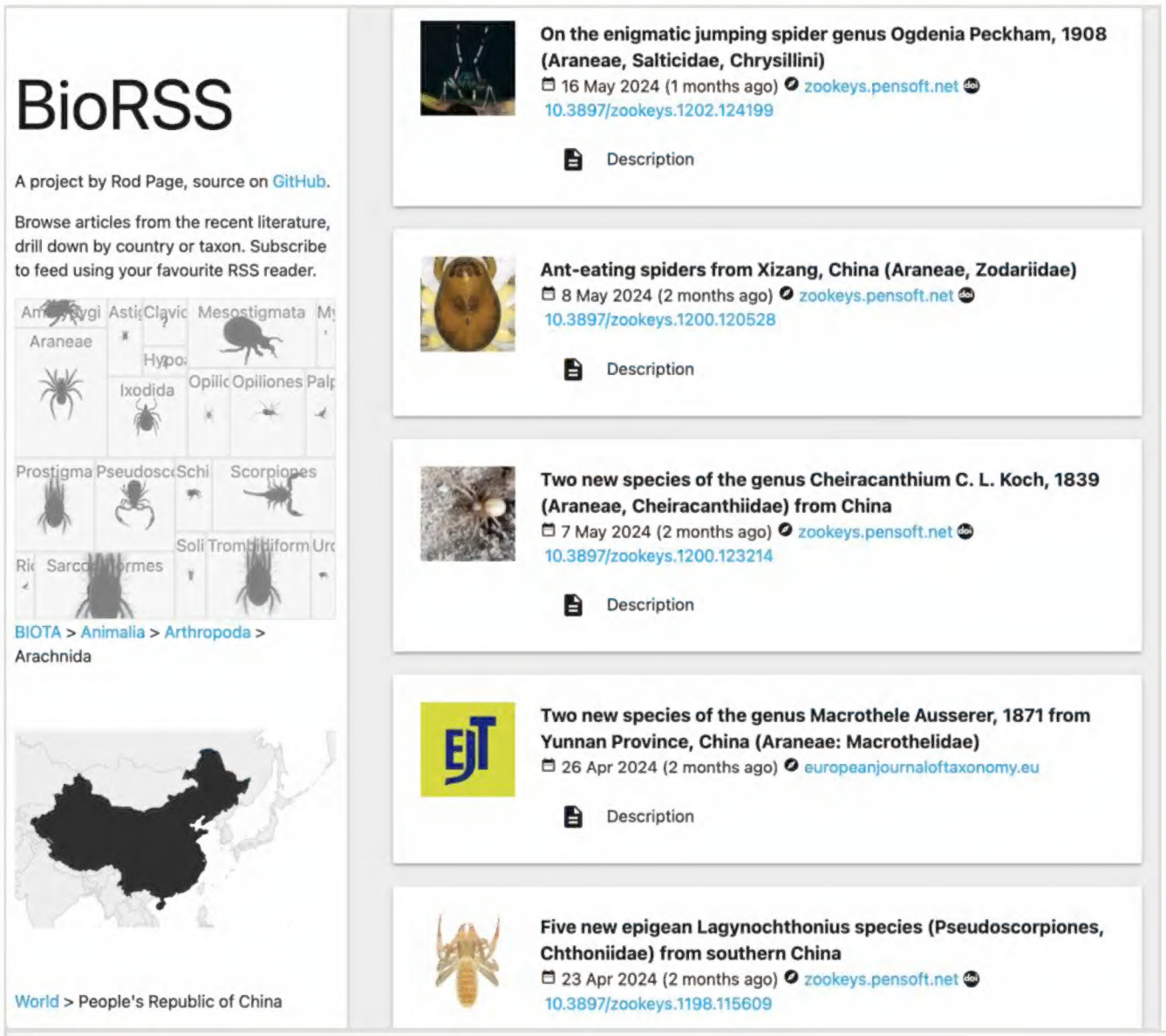
BioRSS
A project by Rod Page, source on GitHub.
Browse articles from the recent literature, drill down by country or taxon. Subscribe to feed using your favourite RSS reader.
Araneae
Ixodida
Opiliones
Scorpiones
BIOTA > Animalia > Arthropoda > Arachnida
World > People's Republic of China
On the enigmatic jumping spider genus Ogdenia Peckham, 1908 (Araneae, Salticidae, Chrysillini)
16 May 2024 (1 months ago) zookeys.pensoft.net
10.3897/zookeys.1202.124199
Description
Ant-eating spiders from Xizang, China (Araneae, Zodariidae)
8 May 2024 (2 months ago) zookeys.pensoft.net
10.3897/zookeys.1200.120528
Description
Two new species of the genus Cheiracanthium C. L. Koch, 1839 (Araneae, Cheiracanthiidae) from China
7 May 2024 (2 months ago) zookeys.pensoft.net
10.3897/zookeys.1200.123214
Description
Two new species of the genus Macrothele Ausserer, 1871 from Yunnan Province, China (Araneae: Macrothelidae)
26 Apr 2024 (2 months ago) europeanjournaloftaxonomy.eu
Description
Five new epigean Lagynochthonius species (Pseudoscorpiones, Chthoniidae) from southern China
23 Apr 2024 (2 months ago) zookeys.pensoft.net
10.3897/zookeys.1198.115609

Figure 1.

Screenshot of BioRSS showing recent papers on Arachnida in China. Other combinations of taxa and geography can be explored using the treemap and geographic maps on the left.

ORCID helpfully provides their data in RDF in JavaScript Object Notation for Linked Data (JSON-LD) format, which we can use to create a simple knowledge graph connecting people, places, publications, and organisations (Fig. 2). ORCID uses the schema.org vocabulary, which simplifies linking together data from disparate sources. Unfortunately, many ORCID profiles lack details on author publications. Even if details on funding and affiliation are included, ORCID lacks information about which works the author published when they had a given affiliation or funding. Data on funding and affiliation for individual publications is, however, often available from CrossRef. Like ORCID, CrossRef supports RDF, but instead of JSON-LD, CrossRef uses XML, and the RDF uses vocabularies such as FOAF (friend of a friend), PRISM (Publishing Requirements for Industry Standard Metadata), and BIBO (Bibliographic Ontology), which for the most part are being superseded by schema.org. Hence for this project I convert metadata from CrossRef into RDF using terms from schema.org (Fig. 3). Works are connected to authors (ideally identified by their ORCID), who in turn are connected to an organisation, ideally with a persistent identifier such as ROR (Research Organization Registry). Works are connected to funders, which may have a DOI from the Open Funder Registry, either directly, or via a grant number.

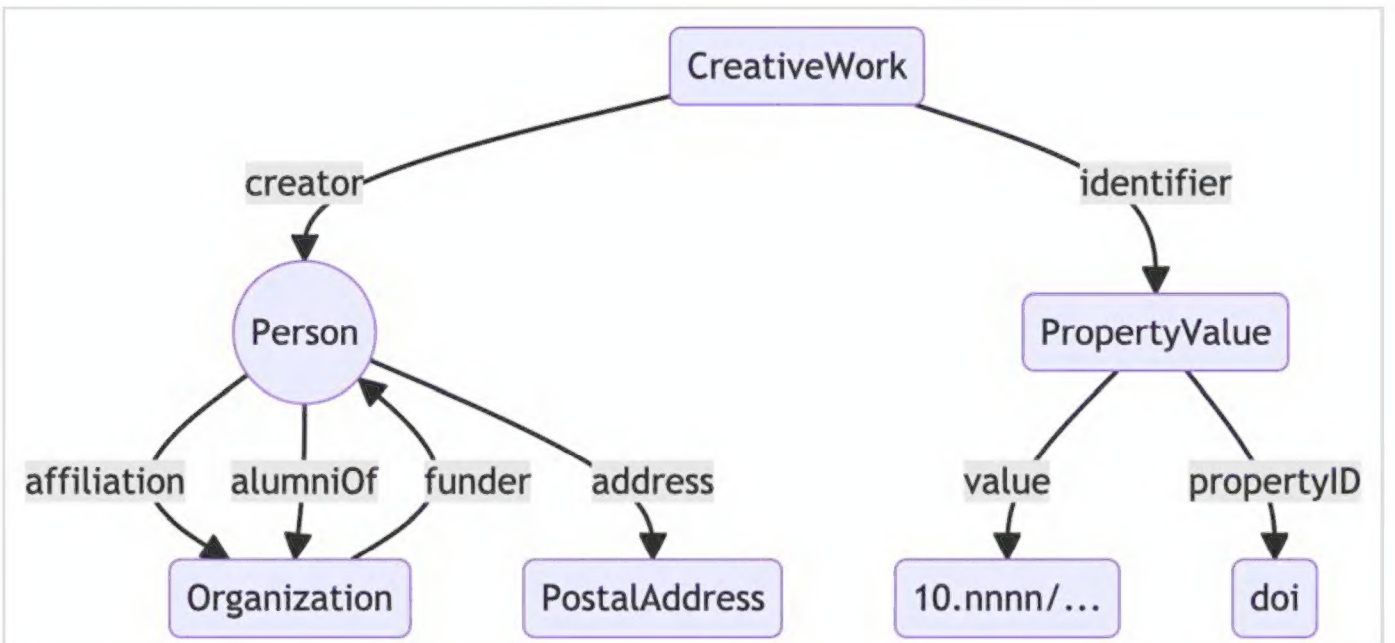

Figure 2.

Simplified version of the data model used by ORCID to export data in RDF. The labels for nodes and edges in the graph come from schema.org.

The final part of the knowledge graph is the connection between taxonomic names and works. One approach would be to use the RSS feeds harvested by BioRSS, which was the original motivation for this work. However, not all the articles BioRSS aggregates are taxonomic, so we would need to be able to reliably filter out non-taxonomic works. In the absence of such a filter I have used lists of recent taxonomic names and publications from Page (2023). Using the DOI for each taxonomic publication, we can connect the taxonomic names to information on authors and funders.

The talk will discuss the construction of this knowledge graph, lessons learnt along the way, and what it tells us about taxonomists and their funders. The talk will also discuss

strategies for the inevitable gap-filling required to flesh out the knowledge graph. Preliminary results reveal that information on author affiliations and funding is often not recorded in either ORCID or CrossRef, which means we will either have to use proprietary databases (such as Dimensions), or scrape it from the Web. The latter approach is likely to benefit from recent developments in machine learning, for example using Large Language Models (LLMs) to parse the acknowledgements section of a paper to extract details on funders and grants. Prospects for these methods will be discussed.

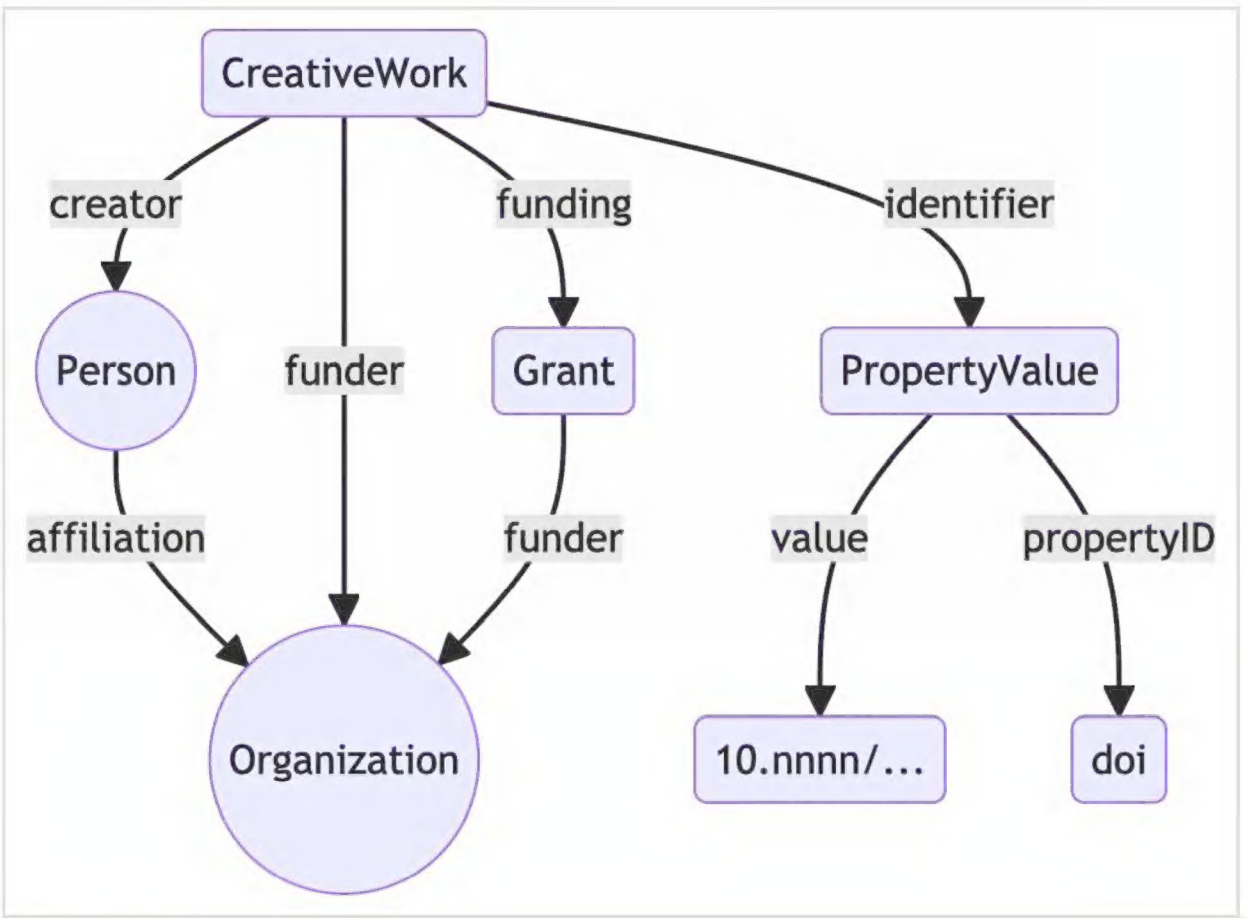

Figure 3.

Simplified data model for a bibliographic record showing links between a work, its author(s) and funder(s). The labels for nodes and edges in the graph come from schema.org.

Keywords

linked data, taxonomy, knowledge graph, funding

Presenting author

Roderic Page

Presented at

SPNHC-TDWG 2024

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Grieneisen M, Zhan Y, Potter D, Zhang M (2014) Biodiversity, Taxonomic Infrastructure, International Collaboration, and New Species Discovery. BioScience 64 (4): 322-332. https://doi.org/10.1093/biosci/biu035
- Joppa L, Roberts D, Pimm S (2011) The population ecology and social behaviour of taxonomists. Trends in Ecology & Evolution 26 (11): 551-553. https://doi.org/10.1016/j.tree.2011.07.010
- Leary P, Remsen D, Norton C, Patterson D, Sarkar IN (2007) uBioRSS: Tracking taxonomic literature using RSS. Bioinformatics 23 (11): 1434-1436. https://doi.org/10.1093/bioinformatics/btm109
- Margush T, McMorris F (1981) Consensus-trees. Bulletin of Mathematical Biology 43 (2): 239-244. https://doi.org/10.1016/s0092-8240(81)90019-7
- Mindell D, Fisher B, Roopnarine P, Eisen J, Mace G, Page RM, Pyle R (2011) Aggregating, Tagging and Integrating Biodiversity Research. PLoS ONE 6 (8). https://doi.org/10.1371/journal.pone.0019491
- Page R (2021) Revisiting RSS to monitor the latest taxonomic research. iPhylo https://doi.org/10.59350/ndtkv-6ve80
- Page R (2023) Ten years and a million links: building a global taxonomic library connecting persistent identifiers for names, publications and people. Biodiversity Data Journal 11 https://doi.org/10.3897/bdj.11.e107914
- Singh V (2017) Replace or Retrieve Keywords In Documents at Scale. arXiv https://doi.org/10.48550/arxiv.1711.00046